Projecting Agricultural Shifts using K-Means Clustering and Gaussian Processes

Mark Lisi
Yale University
New Haven, CT 06511
mark.lisi@yale.edu

Abstract

As the consequences of climate change become more frequent and severe, we are forced to consider the effects of droughts, floods, and storms on one of the world's most foundational industries - agriculture. In this study, we aim to better understand the potential effects of our shifting climate on agriculture using K-means clustering and Gaussian processes. We use PCA and K-means clustering to better understand the relationship between specific crops, and apply a Gaussian process model to historical records of chemical soil composition in and around Kansas to make projections about future shifts.

1 Formulating the Problem

As the climate crisis has continued to intensify in recent years, certain climate-adjacent issues have reached the mainstream; chief among them is global warming. The Intergovernmental Panel on Climate Change reports that human activity has increased average Earth temperature by 1.00C since the industrial revolution, a figure that is likely to increase to 1.50C within the next 20 years [1]. With this trend comes an increase in the frequency and severity of natural disasters: storms, forest fires, droughts, and floods continue to ravage many parts of the world, including the United States. Such disasters have already had significant effects on the world in which we live, and threaten to worsen as time goes on.

In this project, we aim to investigate the ways in which climate change can affect one of society's most foundational industries: agriculture. Although it may be easy to forget in a world of large language models, deep neural networks, and computer vision, we all still live in a heavily agrarian world. Agriculture employs more than 1.3 billion people worldwide, and is responsible for producing more than 80% of the world's food supply. [2] As such, it is paramount to the health of our planet to understand the future of our farms.

2 Predictive Approach

We attempt to better understand the problem of our rapidly changing climate by using machine learning methods to generate reasonable predictions about the future. This sort of prediction can be accomplished first by connecting several relevant, environmental datasets. We will cluster our data to gain a deeper insight into how each parameter works together, focusing on the chemical composition of soil, rainfall, temperature, and anything else that we can expect to change over time. Finally, we can make predictions about the future of the climate and any changes in which crops will be most viable in a given location - we will generate these predictions using a Gaussian process model.

[1] https://www.ipcc.ch/sr15/chapter/spm/

[2] https://www.fao.org/faostat/en/data

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Gaussian processes are well-suited to this task for several reasons. A Gaussian process is a stochastic process, meaning that it functions as a collection of random variables indexed by a continuous variable; space, frequency, and particle size are common choices in other areas, but for our purposes we will index our model on time. By indexing in this way, our model will describe how the system evolves over time. Additionally, a Gaussian process holds the very powerful property that any finite set of its random variables obeys a multivariate normal distribution. This property is particularly useful in the context of prediction since it allows the computation of conditional and predictive distributions. Furthermore, the multivariate normal distribution allows for the use of linear algebra techniques such as Cholesky decomposition, which can significantly speed up computation. Finally, Gaussian processes are non-parametric, meaning that they make no assumptions about the form of the underlying data distribution. This property is particularly useful for complex datasets.

3 Clustering/EDA

3.1 Data Collection

To perform our desired analysis, we collected data that mainly fell into two categories. The first sort of data we needed was the optimal growing conditions for a host of common crops. Within this category, we are able to distinguish between "macro" parameters that externally affect a plant's growth - temperature, rainfall, and humidity, for instance - and "micro" parameters in the chemical composition of soil: nitrogen, phosphorus, potassium, and pH are the elements we focused on in this study. Data of this type was more readily available on public forums such as Kaggle.[3]

The second type of data we needed was more difficult to collect: a historical record for each of these parameters. In order to generate effective predictions, we must sufficiently contextualize our data. Additionally, this category of data should be localized to a specific region to account for general variance across climates. The World Soil Information Service (WoSIS)[4] provides harmonized soil profile data with a massive range in time and region. For our purposes, we have restricted the data to the area between latitudes 36 to 39 and longitudes -103 to -100. This slice is a particularly agrarian chunk of the U.S.A. containing farms mainly in Kansas, but also in Oklahoma, Texas, and Colorado.

3.2 2-dim Plotting

Our first method of exploratory data analysis was plotting our dataset for optimal crop growing conditions in two dimensions, across two parameters at a time. By plotting in this way, we are able to visually understand which crops are sensitive to each variable. We generated two plots in this way: humidity and temperature (Fig. 1) and nitrogen and phosphorus (Fig. 2). In figure 1 we can observe that certain crops that group in vertical strips like papaya, orange, and banana are very insensitive to temperature, but very particular about humidity; conversely, crops like rice can only be grown in a narrow temperature range, but are much more flexible with humidity. In figure 2, crops are much more tightly bracketed - we can conclude that these selected crops are more fickle about the chemical composition of their soil than the external environment.

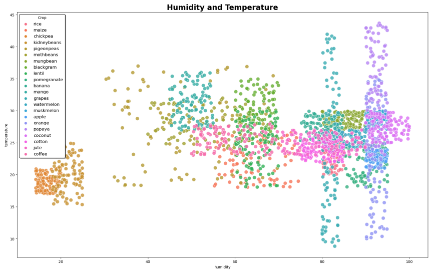

Figure 1: crops indexed over humidity (%) and temperature (C)

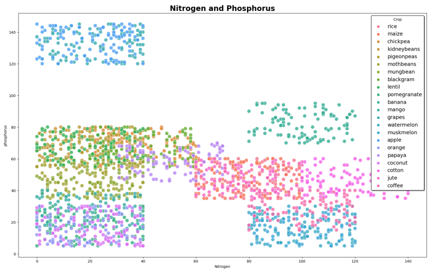

Figure 2: crops indexed over nitrogen and phosphorus content (kg/ha)

[3] https://www.kaggle.com/datasets/aksahaha/crop-recommendation

[4] https://data.isric.org/geonetwork/srv/eng/catalog.search/home

3.3 K-Means Clustering

As a more robust clustering method, we applied K-means clustering to the same dataset. We apply principal component analysis (PCA) to reduce the dimensionality of our dataset to two in order to produce an effective visualization that is easily interpretable, then run K-means clustering with both $k = 3$ and $k = 5$. This type of clustering is valuable because it allows us to visually relate the crops across ALL parameters, and more fully understand which ones are the most similar.

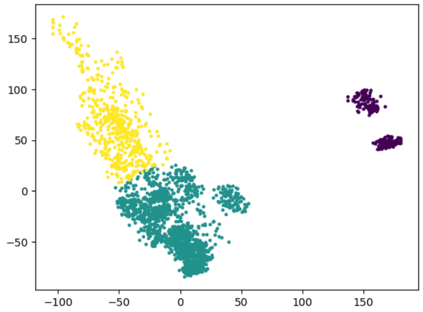

Figure 3: K-means with $k = 3$

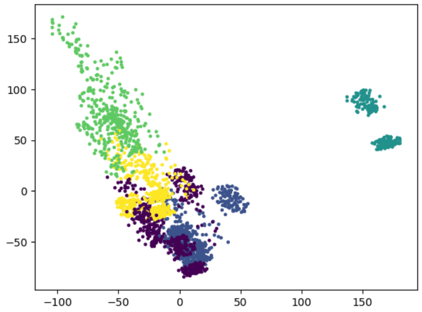

Figure 4: K-means with $k = 5$

The key groupings are the same in the K-means clusters with both $k = 3$ (Fig. 3) and $k = 5$ (Fig. 4) - apples and grapes have their own cluster on the far right, and there is a distinct split in the larger group in on the left that is identified even in the $k = 3$ plot. Figure 4 captures additional information as well - for instance chickpeas are related closely to blackgram and mungbean, but are still somewhat isolated. We include the clustering with the true labels of each point (Fig. 5) to extrapolate further meaning from the K-means plots.

4 GP Findings

As previously mentioned, our Gaussian process is indexed on available temporal data. By indexing this way, we allow the Gaussian process to predict how our system will evolve over time. Since the overall trajectory of our macro-parameters (temperature, etc.) is already widely known, we targeted two different predictive variables with our Gaussian process model: nitrogen and phosphorus. Both nutrients are important for the growth of plants, but nitrogen is highly mobile in soil whereas phosphorus is very immobile. [5] To generate our Gaussian sample, we performed a Cholesky

[5] https://www.noble.org/news/publications/ag-news-and-views/2001/may/mind-your-ps-and-ks/

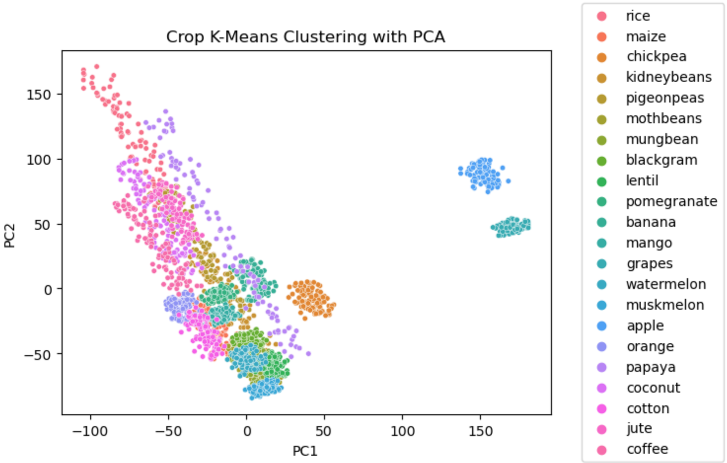

Figure 5: K-means clusters w/ true labels

decomposition over a normal distribution, allowing for more efficient computation. We performed hyperparameter searches for optimal bandwith and noise values.

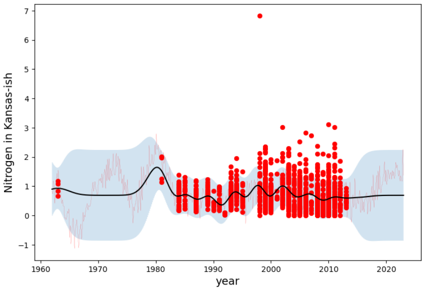

Figure 6: Nitrogen GP Plot

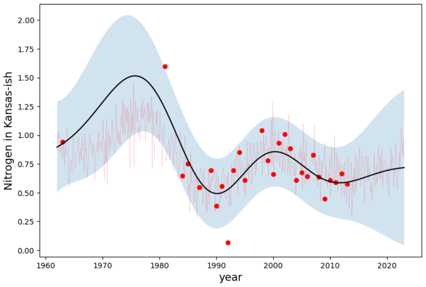

Figure 7: Prediction w/ Yearly Averages

We first ran our Gaussian process model over the levels of nitrogen present in the soil of our selected region over the last 50 years. The plot generated by our model (Fig. 6) highlights some salient properties about our data - there are many measurements per year, and there are multi-year gaps in our dataset where no data is present. The widened uncertainty band reflects the model's increased difficulty in generating a prediction at these times. The posterior mean implies an modest uptick in nitrogen concentration in coming years - with reference to our clustering investigation, this shift could increase the viability of nitrogen-hungry crops like cotton and muskmelon. We then ran the model over the same data averaged by year and observed similar results (Fig 7).

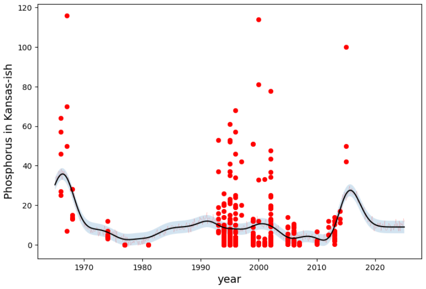

Figure 8: Phosphorus GP Plot

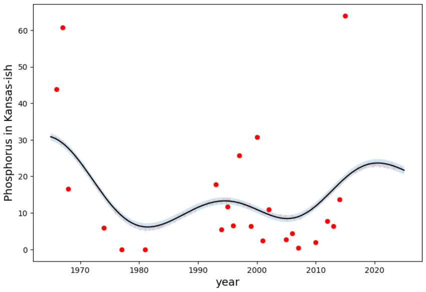

Figure 9: Prediction w/ Yearly Averages

Running our Gaussian process model over historical phosphorus measurements in our selected region yielded less clear results. As the generated plot demonstrates, the WoSIS data in this category was significantly less consistent in both frequency and value; there are long stretches of years without any measurements, and the existing measurements vary dramatically. As a result, our model was not able to functionally predict. One hypothesis as to the cause of this wild variance is the immobility of phosphorus in soil as cited earlier. The sparsity of the data itself is less diagnosable, and is simply a shortcoming of the dataset.

5 Conclusions, Challenges, Future Work

5.1 Conclusions

Based on the results of our Gaussian process model, we can observe a slight upward trajectory in the concentration of nitrogen in the soil between latitudes 36-39 and longitudes -103 to -100 and expect it to continue throughout the 2020s. The measurement values of phosphorus in this region vary dramatically throughout the years, and as a result are much more difficult to predict on. We have already discussed chemical mobility within soil as a potential cause of this variance; in addition, our selected area of the U.S. is extremely inland; such areas are significantly less affected by glacial runoff and coastal erosion, which are two key drivers of climate change.

Through our clustering study, we can observe that certain crops such as papaya, grapes, and oranges are viable within larger temperature ranges than expected. We can draw the same conclusion about pigeonpeas, maize, and jute with respect to the humidity of their growing environment. Still, the impending increase in temperature across the globe may lead society to lean more heavily on more tropical crops. Our clustering also highlights that crops are significantly more sensitive to the chemical composition of their growing environment than any external conditions.

5.2 Challenges

The primary challenge of this study was obtaining sufficient data; for this task, consistent data over a small area is needed. The WoSIS dataset was useful in many ways, but was not consistent enough to generate optimal predictions. When working with such datasets, we encounter a trade-off between ample availability of data and the relevance/consistency of that data. We believe that we struck a satisfactory balance with our region selection, but further scrutiny is always welcome. We also note a high localized variance in the levels of our microparameters: nitrogen, phosphorus, and potassium. This variance is partially attributable to inconsistent data, but is also an inherent property of soil. Lastly, as in any prediction task, mitigating uncertainty is an omnipresent challenge - predicting the future is difficult! While the stochasticity of the Gaussian process model is useful to us in many ways, it may also exacerbate this factor.

5.3 Future Work

As a means of addressing this study's most significant challenge, physical data collection can be carried out! A consistent soil survey in the New Haven County area over multi-decadal timescales would be extremely illuminating in tandem with this study. Additionally, the core ideas of this project can be used with other predictive methods to gain a further understanding of the data. By using robust supervised learning methods such as a multi-layer perceptron or SOTA boosting methods, one could generate a "crop recommender" that takes lat/long coordinates as input and outputs the most viable crops at present day and the expected ones in 5, 10, or 15 years.